多测度的突发词探测及验证研究*

■ 奉国和　武佳佳　莫幸清
华南师范大学经济与管理学院信息管理系　广州 510006

摘　要：[目的/意义]为有效探测科技文献中潜在的研究热点，研究文献中关键词突发的特征条件，构建突发词识别模型对促进科研人员精确把握研究方向具有重要意义。[方法/过程]获取各年度内关键词及词频，构建关键词－年度矩阵，将分析时间段划分为标准窗口、观察窗口和表现窗口，在观察窗口内利用多测度突发词探测模型识别具有突发特征的关键词；在表现窗口内利用LDA挖掘主题词汇作为热点词集合。设计突发词覆盖率指标，辅助滑动时间窗口法，计算不同时间窗口内突发词集合和热点词集合的覆盖率，验证模型识别准确性。[结果/结论]3次滑动时间窗口，计算得到3次突发词覆盖率都在70%以上；与Citespace突发词的对照试验中，本模型3次覆盖率均大于前者，表明设计的突发词探测模型性能良好。

关键词：突发词探测　滑动时间窗口　多测度　LDA主题挖掘
分类号：G250
DOI：10.13266/j.issn.0252－3116.2020.11.008

1　引言

突发词是指词频量较低但增长势头不断增强的关键词，表明该关键词在学科领域受到越来越多的学者关注，未来发展为研究热点概率较大。事物发展遵循基本的生命周期理论，关键词也不例外，在科学传播过程中，关键词发展大致可以分为萌芽期、发展期、成熟期、衰退期4个阶段[1]。关键词作为期刊论文主题、核心概念的集中体现，一定程度上揭示了论文研究内容和研究主题。将关键词作为学科领域突发词探测分析对象，在萌芽期提前识别出具有突发特性的关键词，有利于学者把握学科研究趋势，确定未来研究热点。突发词探测是国内外信息计量学研究领域的重要问题，具有丰富的研究成果，在网络社交媒体中，突发话题探测表现尤为突出。与以往突发词探测研究成果不同，本研究依据科技文献突发词多维度特征，设计突发词探测模型，辅助滑动时间窗口对结果进行验证，并与Citespace突发词探测结果对照。

2　相关研究

目前国内外突发词探测研究方法大体分成三大类：

(1)基于词频增长率进行突发词识别。典型代表是J. Kleinberg提出的突发监测算法(burst detection algorithm，BDA)[2]，该算法认为词的重要性不是词出现的时间长短，而是词出现时的密度，即那些词频相对增长率突然增加的词是突发词[3]。国内外学者基于BDA做了大量研究，并取得了阶段性成果。C. M. Chen[4]基于BDA开发Citespace，对突发词探测进行可视化分析，为科研工作者提供简单易操作的主题探测及演化分析工具[2,5]。唐晓彬等认为Kleinberg使用Viterbi算法仅根据10条是否处于异常状态信息来判断异常事件的发生是不合理的，BDA会将信息频次随时间缓慢变化的状态，误判为有突发异常发生，针对上述缺陷做出BDA改进算法，并成功探测到微博突发事件[6]。卓可秋等认为当文本流无法一次载入内存时，串行计算和多线程单机模式无法在较短的时间内完成突发事件的检测，因此提出MapReduce分布式处理框架解决大数据问题，利用BDA和LDA在新闻数据集得到较好的实验结果[7]。

(2)基于突发词多特征融合进行突发词识别。典型代表是陈国兰采用相对词频、词频增长率和爆发词权重3个指标识别微博文本的突发词，利用共词分析

* 本文系广州市科技计划项目(基础与应用基础研究专题)“突发词探测理论、方法与应用研究”(项目编号：202002030384)研究成果之一。
作者简介：奉国和(ORCID：0000－0002－0774－1544)，教授，博士，E-mail：ghfeng@163.com；武佳佳(ORCID：0000－0002－7342－8388)，硕士研究生；莫幸清(ORCID：000－0001－5481－0349)，硕士研究生。
收稿日期：2019－05－08　修回日期：2019－09－21　本文起止页码：67－76　本文责任编辑：易飞

理论聚类突发词相关事件,成功提取微博突发事件[8];逯万辉等认为单个词语不能表达完整的语义信息,需要从领域术语上探讨该领域知识的演变,因此在构建术语特征词库后,采用频次、频率和词频文档比 3 个指标成功地识别出镍钴产业专利文本中的突发词[9];介飞等认为单一使用文本特征(关键词)或社交行为(评论、点赞、转发)特征都会造成社交网络中隐式突发事件的漏检,将关键词特征和行为特征得到的突发性结果进行关联,有效识别出对比实验中的隐式突发事件[10];W. Xie 等采用 Tweet 总数、词频、词对频次 3 个指标识别 Twitter 中的突发主题,以加速度的计算方式及时反映突发,但该模型可能会忽略短期内不显现突发的主题[11]。

(3)借鉴其他学科理论改进突发词探测方法。王莉亚结合信息熵变化原理,通过观察数据集合加入数据前后的熵值变化判断数据的突发程度,成功解决主题演化发展阶段按照 2 年、5 年或 10 年为单位来划分演化过程是主观且不合理的缺陷[12];王征等认为关键词是科技期刊中承载各类科技概念的最小单位,基于功率谱密度理论和灰色关联理论提出 SRHM 模型,其仿真实验效果好于 Citespace 突发词探测效果,但并未对突发词识别结果做出展示[13];张金柱等认为主题在相邻时间段内的相似度或关联度计算是主题演变及突变识别的核心,而点相似度和关系相似度忽略了网络整体结构,不适于实际网络,因此综合考虑节点数量和重要程度并结合战略坐标图,成功在 WoS 数据集基因编辑领域探测出主题演化进程及突变主题[14];姜鑫等认为小样本数据的关键词词频较低且波动较大,通过计算词频的 Z 分数和移动平均值反映变化趋势并不合适,因此通过对数似然值反映关键词词频变化的显著性程度,由于消除了不同时段科研产出波动对关键词变化趋势的影响,该方法成功识别出科学数据领域基于突发词汇的主题演变过程[15]。随着深度学习技术的广泛应用,有学者开始通过深度神经网络探测突发词,如 L. Shi 等针对微博、Facebook 等社交网络数据提出一个稀疏主题模型(STRM),利用 RNN 学习单词和 IDF 之间的内在关系来测量高频词,模型针对词汇多样性区分突发话题和公共话题[16]。

现有研究成果存在如下几个问题:①突发词探测方法各有不足。第一类方法针对快速流通的数据流如微博具有较好的识别效果,但相比于流通速度较慢的期刊文献不适用;第二类方法,从突发词自身特征出发设置识别条件更具有针对性,可以提高突发词识别精确度,而且适合科技文献按年度划分时间窗的数据类型,但模型设计具有主观性,识别结果不易验证;第三类方法并未在学界广泛使用,其科学性有待验证。②突发词识别对象多为微博短文本、新闻数据、专利文献,对科技文献类数据应用不多。③研究成果并未涉及结果验证,即探测出的突发词是否真为后续的热点。

经过上述分析,本研究借助第二类突发词探测方法基本思想,在相对词频和词频增长率两个通用计量指标基础上,增加词频热度权重指标,反映该关键词在论文标题中出现的频繁程度。其中文献[8]的思想对本研究启发较大,但本研究与其对比有四点明显区别:①研究对象不同:前者使用微博短文本,本文使用科技文献题录信息;②第三个指标选取不同:前者使用 TF-IDF 计算爆发词权重,本文使用词频热度权重计算突发词权重;③研究目的不同:前者使用 k-means 聚类方法识别微博突发事件,本文使用 LDA 挖掘主题词,验证科技文献突发词识别效果,发现科技文献的研究热点;④验证工作不同:前者无验证,本文设计基于时间滑动窗口的验证方法。因此,本文融合上述 3 个特征指标,设计突发词探测模型并提出覆盖率判别指标和滑动窗口方法验证模型效果。

3 构建突发词探测与验证模型

根据前面对突发词本质特性及探测方法分析,设计突发词探测模型。具体步骤如下:

3.1 Step1:获取中频词,构建关键词 - 年度矩阵

词频高低反映其表征的主题特征重要程度,现有成果在确定高频词和高频词选取数量等方面还未达成共识[17],常用的高频词选择方法有两类:一是基于研究者经验;二是结合齐普夫第二定律判断[18]。由于突发词本身的性质决定其词频量不应该为限定时间段内的高频词汇,而通过齐普夫第二定律判断高频词虽然可避免主观性,但设定词频过高,导致中频词范围过大,从而包含部分通用的词汇,此类词汇不属于突发词研究范围。因此,本研究依据图情领域词汇实际应用场景,划分高频、中频和低频词,同时该思想推广到其他领域时,应结合具体的学科特点划分高中低词频。根据上述理论,本研究所指高频关键词是在时间段内总词频大,排在前 N 位的关键词,表明其所体现的主题已被学者广泛传播使用,是处于成熟期的学科基础词汇。低频关键词指在时间段内总词频很小的关键词(本文设置为总词频 1 或 2 的关键词),表明其目前只是被极少数学者关注,未达到广泛关注程度,它不满足

秦国和，武佳佳，莫幸清. 多测度的突发词探测及验证研究[J]. 图书情报工作,2020,64(11):67－76.

本文后续突发词量变需要。中频关键词指在时间段内总词频达到一定量的关键词（本文设置为大于 2 且小于高频关键词频次阈值的关键词）。不同领域的学科基础词汇是不同的，根据学科通识可定义本学科的基础词汇。对比分析发现，中频关键词是最具有研究意义的突发词探测对象，基于中频关键词探测突发词可以避免热点关键词，同时满足突发量变需要。具体来说，本研究定义的高、中、低词频范围，如表 1 所示：

表 1　词频范围划分

词频类型	词频范围（某时间段内）
高频词	词频 > N（N 的取值依据学科常识）
中频词	2 < 词频 < N
低频词	词频 < 3

将关键词按年度采集，统计其词频，并进行同义词、近义词合并以及虚词去除等处理，按照词频大小排序。若某个词连续多年排名靠前，则认为其是专业基础词汇或已是热点词汇，不纳入突发词分析范围。词频年度分布符合长尾分布[19]，词频为 1 或 2 的关键词处于尾部，若多年内某个词汇只出现过 1 或 2 次，则认为该低频词不具备关键词突发特征，同样不纳入突发词分析范围。对排名居中的关键词作进一步筛选，根据二八定律[20]，排名前 20% 的中频词会比剩下的 80% 中频词更具有分析意义，因此将 20% 的中频词作为本文的突发词分析对象，构建关键词－年度矩阵 $F_{m \times n}$。

$$F_{m \times n} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}$$

其中，m 表示关键词数，n 表示年度总数，a_{uv}（$u = 1, \cdots, m; v = 1, \cdots\cdots, n$）表示第 u 个关键词在第 v 年出现的频次。

3.2　Step2：设置分析时间窗口

在关键词－年度词频矩阵中，为验证模型在不同样本矩阵中识别突发词的稳定性，依据时间维度将分析时间段划分为 K 个样本矩阵，定义为 $A_{m \times i}$，$B_{m \times (i+1)}$，$C_{m \times (i+2)}$（$i + 2 < n$）…。为减少时间窗口长度对突发词识别的影响，将每个样本矩阵的时间窗口长度设置一样。同时为保证样本矩阵数据的多样性，设置窗口滑动阈值 T（即每次滑动 T 长度的时间单位），该参数可根据实际需要进行调整，如观察一个单位年度内突发词变化情况可将 T 设置为 1，观察两个或多个单位年度内突发词变化情况可将 T 设置为 2 或大于 2 的数值。为保证突发词变化的时间连续性，本文将 T 设置为 1，即由样本 $A_{m \times i}$ 矩阵开始，滑动一个单位年度可依次得到多个样本矩阵 $B_{m \times (i+1)}$，$C_{m \times (i+2)}$…。同时，将每个样本矩阵划分成 3 个窗口数据矩阵，以样本 $A_{m \times i}$ 矩阵为例，将 i 个单位年度划分为 3 个时间窗口数据。时间在前的窗口为标准窗口，时间居中的窗口为突发词探测窗口，称为观察窗口；时间在后的窗口为热点主题探测窗口，称为表现窗口。将上述 3 个窗口用符号分别定义为 AT_1，AT_2，AT_3。为满足关键词突发量变需要及可计算性，设置 AT_1，AT_2，AT_3 的窗口长度为 3 个单位年度。标准窗口数据是关键词突发变化的比对标准，对观察窗口内的数据依据突发词特征条件进行判断，满足条件的关键词归入突发词集合；将表现窗口所有频次的词汇通过 LDA[21] 分析挖掘热点主题，并设置阈值选择每个主题内概率值排在 TopN 关键词归入热点词集合。

以固定窗口大小滑动一个单位年度，获取 $A_{m \times i}$，$B_{m \times (i+1)}$，$C_{m \times (i+2)}$（$i + 2 < n$），3 个样本矩阵的覆盖率，分别表示为 p_A，p_B，p_C。本文设计的滑动窗口及分析矩阵如图 1 所示：

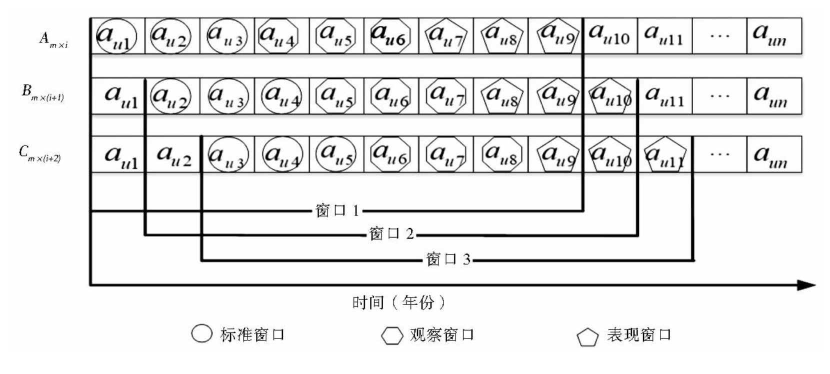


图 1　分析时间窗口划分

3.3 Step3：探测突发词

综合考虑现有文献的突发词特征指标，剔除针对微博短文的话题标签指标[22]、TF-IDF 权重指标[23]，设计并约定突发词应该在观察窗口内并满足以下基本条件：①量变条件：关键词总词频要达到一定的量，导致质变而引起后续突发；②趋势条件：关键词的词频逐年增多，呈上升趋势；③波动性条件：关键词词频波动大，区分度强。由于期刊文献的更新周期较长，单位年度内无法达到上述条件①的标准，所以对 3 个时间窗内的关键词词频分别求总和。依据上述基本条件设置如下启发式描述量：

（1）相对词频。计算关键词词频和当前窗口内最大关键词词频的比率，如公式（1）所示：

$$X = \frac{\sum a_{mn}}{\max\left(\sum a_{mn}\right)} \qquad \text{公式(1)}$$

其中 X 表示关键词 M 在 n 年内的相对词频，max（$\sum a_{mn}$）表示 n 年内词频最大的值。X 考察关键词在垂直方向的变化趋势，X 值越大说明该关键词在时间窗内热度越大，在未来可能会成为热点词汇。以样本 $A_{m\times t}$ 矩阵为例，在对应的 AT_1，AT_2，AT_3 时间窗口内，通过公式（1）计算会得到相应的相对词频，X_{AT_1}，X_{AT_2}，X_{AT_3}。

（2）词频增长率。计算当前窗口内的词频相对前一个窗口增长的比率，如公式（2）所示：

$$Z = \frac{\sum a_{mn} - \sum a_{m(n-1)}}{1 + \sum a_{m(n-1)}} \qquad \text{公式(2)}$$

其中，Z 表示关键词 M 在 n 年内的词频相对于 $n-1$ 年内的增长率，$1+\sum a_{m(n-1)}$ 可以避免前一个时间段内，关键词未出现导致分母为 0 的情况。Z 考察关键词在水平方向上的变化趋势，Z 值越大说明该关键词的增长趋势越明显，越有可能会成为热点词汇。

（3）词频热度权重。计算科技文献题目中包含关键词的比率，如公式（3）所示：

$$H = \frac{\sum t\{a_{mn}\}}{\sum title} \qquad \text{公式(3)}$$

其中，H 表示关键词 M 出现在题目中的数量与当前时间内总题目数量的比率，$\sum t\{a_{mn}\}$ 表示题目中包含关键词的数量，$\sum title$ 表示文献题目总条数。H 值越大说明该关键词在题目中出现的次数越多，该词在当前时间窗口内热度越大，未来越有可能成为热点词汇。

（4）将依据描述量 X，Z，H 筛选出的关键词分别归入突发词候选集合 X_{ht}，Z_{ht}，H_{ht}。设置阈值 s，将各集合内按照描述量排名位次大于 s 的关键词归入突发词集合 T。用数学符号表示即：

$$T = \{X_{ht} \cap Z_{ht} \cap H_{ht}\}(\text{描述量} > s) \qquad \text{公式(4)}$$

3.4 Step4：捕获热点词

热点词是表现窗口内频次高且稳定的关键词，热点词获取范围应该大于突发词分析范围，以保证突发词在后续时间窗口成为热点词的可能性。LDA 语言模型是一种基于三层贝叶斯概率模型[24]，包含词、主题和文档 3 层结构，是目前比较成熟的文档主题生成模型，与共词分析挖掘热点相比，LDA 具有三大优势[1]：①无需确定高频低频关键词分界线；②LDA 可反映主题词之间深层次的语义关系；③避免共词分析关键词选择的主观性。利用该模型挖掘出表现窗口内热点主题及各主题包含的关键词，即本文需要的热点词。

定义文档集合符号为 $D=\{d_1,\cdots,d_p\}$，d_p 表示第 p 篇文档，$d_p=\{x_1,\cdots,x_j\}$，x_j 表示第 p 篇文档中第 j 个词汇。主题符号为 $E=\{k_1,\cdots,k_o\}$，k_o 表示主题内第 o 个关键词。热点关键词的计算公式如下：

$$\mathrm{P}(k \mid d) = \mathrm{P}(k \mid e) * \mathrm{P}(e \mid d) \qquad \text{公式(5)}$$

其中，k，d，e 分别表示关键词、文档、主题。依据公式（5），得到文档集合中主题以及每个主题包含的关键词。设置 LDA 超参数 q 调整 LDA 生成的主题数目，将小于 q 的主题词汇归入热点词集合，定义热点词集合为 R。

3.5 Step5：模型验证

为验证模型识别突发词效果，提出覆盖率判别指标，即选择突发词集合和热点词集合的共同词汇，并计算相同词汇占突发词集合的比率，其定义如下：

$$P = \frac{T \cap R}{T} \qquad \text{公式(6)}$$

其中，P 是覆盖率，覆盖率越大，表示观察窗口得到的突发词与表现窗口得到的热点词对应程度越高，模型性能越好。T，R 即由 3.3、3.4 得到的突发词集合和热点词集合。

为保证模型在不同时段样本的适用性，采用滑动窗口的方法，将标准窗口、观察窗口与表现窗口往后移动一个单元（即移动 1 年），保持 3 个窗口长度不变，依据上述步骤重复计算覆盖率，依次得到 p_A，p_B，p_C，根据不同样本的覆盖率判断突发词探测模型的稳定性。

4 实证分析

将模型思想应用于图情领域科技文献突发词探测，在 CNKI 上采集 2007－2017 年间 18 种 CSSCI 图情

核心期刊的文献信息，每条数据结构如下：{作者，题名，关键词，年份}。按照图 1 分析时间窗口的划分方法，n 的长度为 11，滑动窗口间隔为 1，$A_{m\times i}$，$B_{m\times(i+1)}$，$C_{m\times(i+2)}$（$i+2<n$）3 个样本矩阵的时间长度为 9。每个样本矩阵的标准窗口、观察窗口、表现窗口的长度均为 3。2007 – 2017 年分析时间窗口划分如表 2 所示：

表 2　分析时间窗口划分

数据样本	标准窗口	观察窗口	表现窗口	覆盖率
$A_{m\times i}$	AT_1	AT_2	AT_3	p_A
$B_{m\times(i+1)}$	BT_1	BT_2	BT_3	p_B
$C_{m\times(i+2)}$	CT_1	CT_2	CT_3	p_C

4.1　数据预处理

初始数据结构：{题名，作者，关键词，年份}是由 4 个元素组成，共计 53 221 条记录。其中需要处理的记录如表 3 所示，主要包括以下 3 类：①数据缺失：没有关键词、题名、作者等信息的记录；②非期刊论文：题名中包含举办、举行、委员会、讲话、致辞等词汇的征文、通告；③特殊字符：如“；；”和“；”。使用 python 的 pandas 和 numpy 工具做数据分析，使用 excel 和 sqlite 做数据保存工具。

表 3　待处理的数据举例

题名	作者	关键词	年份
投稿指南		文后参考文献；稿件类型；文献序号	2017
深阅读：概念构建与路径探索	李桂华；	阅读推广；；深阅读；；阅读参与；；阅读行为	2017
…	…	…	…

按照上述情况，删除①类和②类错误，整理③类数据。同时建立同义词表和停用词表，合并意义相近、英语大小写、中英同义的关键词，将“先生”“特点”“文章”等没有研究意义的词汇归入停用词表，并在关键词表中剔除。

4.2　构建关键词 – 年度词频矩阵

根据模型，首先构建 $F_{m\times n}$，以关键词列为唯一索引，以年度为列名，关键词词频为矩阵元素值。因为要满足关键词量变的基础，所以剔除词频低于 3 个以下的关键词，同时剔除 11 年内一直排在词频前部的学科基础词汇，如图书馆、高校图书馆、公共图书馆等。在此基础上依据二八定律，构建维度是 1904 * 11 的 $F_{m\times n}$。所选关键词占总词汇的 24%，符合二八定律。整理完成后的 $F_{m\times n}$ 如表 4 所示（注：0717 总词频表示某词汇在 2007 年到 2017 年间的词频和）。

表 4　关键词 – 年度词频矩阵

序号	关键词	2007 年词频	…	2017 年词频	0717 总词频
1	信息服务	77	…	31	848
2	知识管理	94	…	20	751
3	竞争情报	50	…	21	608
…	…	…	…	…	
1904	链路预测	0	…	4	10

依据 3.2，将表 4 的 $F_{m\times n}$ 矩阵按照划分好的时间窗口切割成 3 个样本矩阵，即上述 $A_{m\times i}$，$B_{m\times(i+1)}$，$C_{m\times(i+2)}$（$i+2<n$）。因为各样本的时间窗口及计算方法一致，所以本文以样本 $A_{m\times i}$ 矩阵为例进行突发词探测及验证。

4.3　捕获突发词

依据 3.3，合并样本 $A_{m\times i}$ 矩阵中每 3 年的词频总和，即 $\sum a_{mn}$。参照公式(1)、公式(2)、公式(3)分别计算出 AT_2 的相对词频，词频增长率及词频热度权重。经过实验，当 $s=200$ 时，模型效果较好，即将 AT_2 窗口内各指标排名前 200 的关键词纳入突发词候选集合。参照公式(4)，计算 3 个集合的交集，共得出 13 个突发词。突发词结果如表 5 所示：

表 5　样本 $A_{m\times i}$ 矩阵 AT_2 窗口相对于 AT_1 窗口的突发词

关键词	词频/1 年									词频总和/3 年			AT_2 突发词指标		
	07	08	09	10	11	12	13	14	15	AT_1	AT_2	AT_3	X	Z	H
微博				5	15	51	65	54	47	0	71	166	0.51	71.00	7.51
关联数据				3	15	30	33	35	39	0	48	107	0.35	48.00	3.12
云计算			13	35	45	58	46	43	25	13	138	114	1.00	8.93	6.69
突发事件		1	2	13	12	13	14	19	30	3	38	63	0.28	8.75	3.87
知识图谱		2	6	19	27	39	41	40	36	8	85	117	0.62	8.56	4.83
学科服务	1	4	6	29	31	53	59	41	36	11	113	136	0.82	8.50	7.21
网络舆情			8	20	26	27	39	67	60	8	73	166	0.53	7.22	5.80
阅读推广	1	2	3	6	21	28	42	44	73	6	55	159	0.40	7.00	4.24
研究热点	2	1	6	17	23	25	22	18	13	9	65	53	0.47	5.60	5.50
信息行为	5	3	3	14	21	21	24	31	27	11	56	82	0.41	3.75	3.50
服务体系		5	3	14	12	15	14	11	7	8	41	32	0.30	3.67	7.88
虚拟社区	2	3	3	12	15	11	19	10	16	8	38	45	0.28	3.33	2.90
文献计量分析	1	3	4	10	18	10	6	4	7	8	38	17	0.28	3.33	3.64

注：07 即 2007 年，其后各年依此类推；AT_1，AT_2，AT_3 即步骤 2 的标准窗口、观察窗口、表现窗口；X，Z，H 即步骤 3 的相对词频、词频增长率、词频热度权重

4.4　捕获热点词

依据模型步骤 4，将原始数据{题名，作者，关键词，年份}中关键词词条列作为 LDA 挖掘语料，关键词词条列即步骤 4 的 D 文档集合，词条中每个关键词即 $d_p = \{x_1, \cdots, x_j\}$ 每篇文档中的关键词。将 AT_3 窗口内每篇期刊文献的关键词词条分词后，去除文档集合中没有实际意义的停用词，作为 LDA 模型的输入文档集合。利用 gensim 文本分析工具对文本集进行训练，经过多次实验，发现 $q = 10$ 时，模型效果较好，即设置主题数目为 10，每个主题包含概率值排前 10 的关键词。AT_3 窗口内的主题词汇如表 6 所示：

表 6　样本 $A_{m\times i}$矩阵 AT_3 窗口内的热点词集合

主题 1	主题 2	主题 3	主题 4	主题 5	主题 6	主题 7	主题 8	主题 9	主题 10
知识服务	公共图书馆	图书馆	可视化	数字图书馆	社会网络分析	知识管理	本体	高校图书馆	***学科服务***
竞争情报	图书情报学	数据库	电子政务	信息服务	大学图书馆	知识共享	***阅读推广***	***网络舆情***	信息资源
情报学	知识组织	用户需求	专利分析	图书馆服务	数据挖掘	***虚拟社区***	移动图书馆	图书馆学	引文分析
云计算	图书馆员	共建共享	开放获取	***知识图谱***	文本挖掘	元数据	信息素养	***微博***	学科馆员
大数据	比较研究	大专院校	资源建设	社会网络分析	聚类分析	***文献计量学***	研究进展	政府信息资源	服务模式
图书馆联盟	图书馆事业	文献传递	知识转移	开放存取	共词分析	信息检索	电子书	嵌入式服务	移动服务
文献计量	网络社区	科技查新	知识库	***信息行为***	统计分析	***突发事件***	期刊评价	信息共享空间	学术影响力
关联数据	信息技术	服务创新	机构知识库	评价指标	信息分析	搜索引擎	手机图书馆	参考咨询	国际合作
资源共享	网络结构	版权	***研究热点***	信息需求	用户行为	知识创新	信息生态	信息传播	馆藏建设
数字资源	研究综述	服务质量	读者服务	专利	信息素养教育	结构方程模型	知识地图	学科化服务	专利地图

注：加粗斜黑体字表示突发词与热点词的重叠

由表 6 发现，每个主题均由 10 个关键词构成。其中，主题 1 描述云计算和大数据在图书情报领域内提供资源共享和知识服务的功能；主题 2 描述公共图书馆的知识组织和图书馆馆员及图书馆的事业发展；主题 3 描述图书馆服务，包括文献传递、科技查新及服务质量；主题 4 描述信息资源的建设，包括电子政务、机构知识库及资源的开放获取；主题 5 描述信息服务，包括图书馆服务、知识图谱、用户行为及用户需求；主题 6 描述图书情报学科的分析方法，包括数据挖掘、文本挖掘、聚类分析、共词分析；主题 7 描述知识共享社区，包括知识管理、知识共享、虚拟社区及知识创新；主题 8 描述图书馆阅读推广活动，以此提高读者的信息素养；主题 9 描述网络舆情的发展，包括舆情产生的传播工具——微博及图书馆服务；主题 10 描述图书情报领域的学科服务，包括学科馆员素质的提升、服务模式的改进以及学术影响力。

综上所述，表现窗口关键词涉及的主要主题为图书馆服务（科技查新、文献传递、阅读推广），新兴技术（大数据、云计算、知识图谱），学科方法（数据挖掘、文本挖掘、共词分析、社会网络分析）等，此外还有专利分析、信息素养、用户需求、知识库。

4.5　计算突发词覆盖率

依据模型步骤 5，将 4.3 筛选出的突发词集合与 4.4 筛选出的热点主题词进行突发词覆盖率计算。参照公式（6），$P = \frac{T \cap R}{T} = \frac{12}{13} = 0.92$。该结果表明在样本 $A_{m\times i}$矩阵中由突发词探测模型在 AT_2 窗口识别的突发词，有 92% 的准确率表现在 AT_3 窗口内。同时，突发词集合 = {*微博，关联数据，云计算，突发事件，知识图谱，学科服务，网络舆情，阅读推广，研究热点，信息行为，服务体系，虚拟社区，文献计量学*}较全面地反映了热点主题词（集合 T 与集合 R 交集所包含的元素，即表 6 中斜黑体字）。

4.6　滑动窗口分析

依据分析时间窗口的设置，重复 4.1－4.5 的计算步骤，依次可得样本 $B_{m\times(i+1)}$矩阵、样本 $C_{m\times(i+2)}$矩阵的 p_B，p_C。

（1）样本 $B_{m\times(i+1)}$矩阵中 窗口突发词集合 = {阅读推广，移动图书馆，微博，社会网络分析，免费开放，可视化分析，科学数据，关联数据，大数据}，共计 9 个突发词。样本 $B_{m\times(i+1)}$矩阵 窗口的突发词结果如表 7 所示，样本 $B_{m\times(i+1)}$矩阵中 BT_3 窗口的主题词汇如表 8 所示。

由表 8 可以看出，主题 1 描述情报学科的分析方法；主题 2 描述信息资源开放获取及知识组织的常用形式，包括知识库、本体、元数据等；主题 3 描述图书馆服务内容及方式，包括馆藏服务、推荐服务、电子资源管理等；主题 4 描述大数据技术在现代信息网络的应用；主题 5 描述图书馆数据库管理系统及图情领域行

表 7　样本 $B_{m\times(i+1)}$ 矩阵 BT_2 窗口相对于 BT_1 窗口的突发词

关键词	词频/1 年									词频总和/3 年			BT_2 突发词指标		
	08	09	10	11	12	13	14	15	16	BT_1	BT_2	BT_3	X	Z	H
大数据					9	42	68	104	140	0	51	312	0.39	51	3.3
微博			5	15	51	65	54	47	40	5	131	141	1	21	14.48
关联数据			3	15	30	33	35	39	45	3	78	119	0.6	18.75	4.9
免费开放	1	1		20	11	10	4	1		2	41	5	0.31	13	2.76
移动图书馆	1	3	4	18	22	42	40	30	33	8	82	103	0.63	8.22	3.76
阅读推广	2	3	6	21	28	42	44	73	82	11	91	199	0.69	6.67	7.2
可视化分析	1		1	6	7	9	9	11	9	2	22	29	0.17	6.67	4.44
科学数据	2	2	2	7	8	26	27	28	26	6	41	81	0.31	5	3.68
社会网络分析	3	9	18	42	42	43	51	42	32	30	127	125	0.97	3.13	2.84

注:08 即 2008 年,其后各年依此类推;BT_1,BT_2,BT_3 分别对应步骤 2 的标准窗口、观察窗口、表现窗口;X,Z,H 分别对应步骤 3 的相对词频、词频增长率、词频热度权重

表 8　样本 $B_{m\times(i+1)}$ 矩阵 BT_3 窗口内的热点词集合

主题 1	主题 2	主题 3	主题 4	主题 5	主题 6	主题 7	主题 8	主题 9	主题 10
竞争情报	开放获取	图书馆联盟	***大数据***	图书馆	图书馆服务	高校图书馆	公共图书馆	数据共享	图书馆学
情报学	云计算	服务质量	***微博***	数据库	信息服务	数字图书馆	网络舆情	企业管理	学科馆员
阅读推广	本体	馆藏	***社会网络分析***	版权	高校图书馆	知识服务	情报学	***科学数据***	电子
知识图谱	机构知识库	通检	信息检索	数据库系统	信息素养	图书馆员	开放存取	图书情报学	轻子
可视化	***移动图书馆***	个性化推荐	社会网络	微信	文献资源建设	***关联数据***	文献计量学	科研数据	服务模式
情报工作	元数据	电子资源管理	聚类分析	信息管理	电子书	学科服务	***突发事件***	个性化服务	嵌入式学科服务
知识共享	知识组织	研究综述	信息服务	比较研究	数据挖掘	学科馆员制度	***微博***	用户需求	数字资源
专利分析	信息行为	馆藏建设	评价指标	NISO	指标体系	图书馆工作人员	需求驱动	信息资源	专利
文献计量	检索工具	知识转移	专著	EBSCO	层次分析法	知识管理	学术思想	科研数据管理	信息需求
共词分析	阅读推广活动	网络信息资源	OCLC	IMLS	ODI	移动服务	ProQuest	Web2.0	SirsiDynix

注:突发词集合以加粗斜黑体字表示

业协会;主题 6 描述高校图书馆的信息服务及读者信息素养;主题 7 描述图书馆馆员及学科服务主题;主题 8 描述网络信息资源的发展,包括网络舆情、微博、情报学、突发事件及数据资源的开放存取;主题 9 描述数据共享主题,包括科研数据管理和企业数据管理;主题 10 描述图书馆学与其他学科的交叉性,包括电子、轻子等理工类学科。参照公式(6),$P = \frac{T\cap R}{T} = \frac{8}{9} = 0.89$。结果表明在样本 $B_{m\times(i+1)}$ 矩阵中由突发词探测模型在 BT_2 窗口识别的突发词,有 89% 的准确率表现在 BT_3 窗口内。

(2)样本 $C_{m\times(i+2)}$ 矩阵中 CT_2 窗口突发词集合 = {*云服务,阅读推广,移动图书馆,微博,微信,数据管理,馆藏资源,科学数据,关联数据,大数据*},共计 10 个突发词。样本 $C_{m\times(i+2)}$ 矩阵 CT_2 窗口的突发词结果如表 9 所示,样本 $C_{m\times(i+2)}$ 矩阵中 CT_3 窗口的主题词汇如表 10 所示。

由表 10 可以看出,主题 1 描述信息组织和信息分析的方法,包括元数据、协同过滤、聚类分析;主题 2 描述云计算技术在图情领域的应用及数据的可视化;主题 3 描述高校图书馆馆员信息素养的提升及阅读推广活动,新兴在线学习空间——慕课;主题 4 描述图书馆使用关联数据、数字化等技术建设馆藏资源;主题 5 描述图情领域新方法,包括知识图谱、社会网络分析;主题 6 描述数据开放获取运动,强调信息知识化、知识共享化,包括知识组织、知识产权、机构知识库;主题 7 描述图书情报学的基本工作,包括数据治理技术、数据组织技术、数据表示技术;主题 8 描述内容不明显,涉及多个主题词汇,包括专利分析、移动图书馆等;主题 9 描述网络舆情的发展,包括网络突发事件、数据挖掘技术等;主题 10 描述图书馆阅读推广活动、数字阅读、全民阅读等服务以及数据共享和数据开放。参照公式(6),$P = \frac{T\cap R}{T} = \frac{7}{10} = 0.7$。结果表明在样本 $C_{m\times(i+2)}$ 矩阵中由突发词探测模型在 CT_2 窗口识别的突发词,有 70% 的准确率表现在 CT_3 窗口内。

表 9　样本 $C_{m\times(i+2)}$ 矩阵 CT_2 窗口相对于 CT_1 窗口的突发词

关键词	词频/1 年									词频总和/3 年			CT_2 突发指标		
	09	10	11	12	13	14	15	16	17	CT_1	CT_2	CT_3	X	Z	H
大数据				9	42	68	104	140	114	0	119	358	0.70	119.00	9.09
微信					7	21	24	33	27	0	28	84	0.16	28.00	3.44
微博		5	15	51	65	54	47	40	34	20	170	121	1.00	7.14	22.36
数据管理		1	2	4	9	12	12	22	24	3	25	58	0.15	5.50	3.93
关联数据		3	15	30	33	35	39	45	31	18	98	115	0.58	4.21	5.82
科学数据	2	2	7	8	26	27	28	26	24	11	61	78	0.36	4.17	5.65
云服务	2	1	5	13	17	9	7	4	5	8	39	16	0.23	3.44	2.70
馆藏资源	1	4	1	9	5	14	8	5	1	6	28	14	0.16	3.14	2.62
移动图书馆	3	4	18	22	42	40	30	33	33	25	104	96	0.61	3.04	6.14
阅读推广	3	6	21	28	42	44	73	82	81	30	114	236	0.67	2.71	9.09

注:09 即 2009 年,其后各年依此类推; CT_1,CT_2,CT_3 即步骤 2 的标准窗口,观察窗口,表现窗口;X,Z,H 即步骤 3 的相对词频,词频增长率,词频热度权重

表 10　样本 $C_{m\times(i+2)}$ 矩阵 CT_3 窗口内的热点词集合

主题 1	主题 2	主题 3	主题 4	主题 5	主题 6	主题 7	主题 8	主题 9	主题 10
知识服务	可视化	高校图书馆	图书馆	知识图谱	开放获取	***大数据***	专利分析	网络舆情	公共图书馆
元数据	云计算	竞争情报	信息服务	社会网络分析	大学图书馆	情报学	***移动图书馆***	突发事件	***阅读推广***
文献计量	评价指标	信息素养	数字图书馆	引文分析	学科服务	本体	知识管理	数据挖掘	微博
虚拟社区	出版物	图书馆员	图书馆学	共词分析	机构知识库	图书情报学	信息行为	信息检索	数字资源
指标体系	图书馆联盟	阅读推广	***关联数据***	服务模式	公共文化服务	语义网	社交网络	信息组织	全民阅读
知识转移	版权	***微信***	图书馆服务	社会网络	比较研究	慕课	信息安全	专利分析	***科学数据***
聚类分析	文本分类	信息素养教育	资源建设	用户行为	知识产权	法人治理结构	文献计量学	文本挖掘	情报分析
协同过滤	基层图书馆	结构方程模型	服务创新	著作权	社交媒体	公共文化	知识共享	智库	数字阅读
移动互联网	社会网络分析	慕课	电子书	研究热点	知识组织	情报研究	社会化媒体	评价体系	数据共享
EBSCO	翻转课堂	Springer	PreQuest	信息需求	智慧图书馆	情报工作	知识网络	个人信息	开放数据

注:突发词集合以加粗斜黑字体表示

4.7　对照实验

为验证新模型性能,采用主流突发词探测工具 Citespace 作对照实验,数据源和突发词探测时间段同上。在 Citespace 软件选择 Burstness 检测,参数设置如下:将每年词频大于 50 的词汇作为候选突发词集,即 Select Top = 50;在 Burstness 面板选择词汇最低突发持续时间为 1 年,即 Minimum Duration = 1,根据突发强度值排名得到不同数据样本的突发结果,2010 – 2012 年、2011 – 2013 年、2012 – 2014 年依次对应表 11、表 12、表 13。依据时间变化趋势发现,Citespace 探测到的突发词包含消亡趋势(如:信息素质、个性化服务)和上升趋势(如:文献计量学、共词分析)两种类型。本研究认为上升型突发词在未来更有可能成为研究热点,对学科研究方向更具有指导意义,因此新模型更注重具有上升趋势的突发词汇。

表 11　Citespace 在 2010 – 2012 年探测出的突发词

突发词	年份	强度值	开始年份	结束年份	2010 – 2012
文献计量学	2010	12.499 5	2011	2012	
共词分析	2010	12.118 7	2011	2012	
统计分析	2010	12.118 7	2011	2012	
聚类分析	2010	11.738 1	2011	2012	
信息素质	2010	11.736 2	2010	2010	
个性化服务	2010	11.080 5	2010	2010	
知识组织	2010	10.752 9	2010	2010	
信息需求	2010	10.596 9	2011	2012	
信息资源共享	2010	9.836 7	2011	2012	
评价指标	2010	9.770 6	2010	2010	
服务质量	2010	9.770 6	2010	2010	
图书馆事业	2010	9.443 4	2010	2010	
电子资源	2010	9.116 3	2010	2010	
知识创新	2010	8.462 4	2010	2010	
图书馆管理	2010	6.352 8	2010	2010	
社会网络分析	2010	5.837 8	2011	2012	
专利分析	2010	3.313 5	2011	2012	
隐性知识	2010	2.175 7	2010	2010	
著作权	2010	1.803 0	2010	2010	
开放存取	2010	1.606 8	2011	2012	

注:加粗黑体字为 Citespace 突发词与热点词集合重叠的词汇,表 12、表 13 同

表 12　Citespace 在 2011 – 2013 年探测出的突发词

突发词	年份	强度值	开始年份	结束年份	2011 – 2013
信息共享空间	2011	12.274 6	2011	2011	
统计分析	2011	11.548 6	2011	2011	
知识产权	2011	11.185 7	2011	2011	
聚类分析	2011	11.185 7	2011	2011	
信息需求	2011	10.098 0	2011	2011	
电子商务	2011	9.735 7	2011	2011	
信息资源共享	2011	9.373 5	2011	2011	
图书馆管理	2011	9.373 5	2011	2011	
知识转移	2011	9.113 4	2011	2011	
著作权	2011	9.061 2	2012	2013	
比较研究	2011	9.061 2	2012	2013	
隐性知识	2011	8.697 2	2012	2013	
元数据	2011	8.333 4	2012	2013	
web2.0	2011	6.949 1	2011	2011	
读者服务	2011	4.657 9	2011	2011	
信息检索	2011	2.993 4	2011	2011	
开放存取	2011	2.604 5	2011	2011	
机构知识库	2011	2.402 5	2011	2011	
共词分析	2011	2.140 6	2011	2011	
图书馆员	2011	1.891 7	2011	2013	
文献计量学	2011	1.827 7	2011	2011	
社会网络	2011	1.753 3	2012	2013	
科技查新	2011	1.702 7	2011	2011	

表 13　Citespace 在 2012 – 2014 年探测出的突发词

突发词	年份	强度值	开始年份	结束年份	2012 – 2014
web2.0	2012	11.262 9	2012	2012	
知识转移	2012	10.897 7	2012	2012	
文献计量学	2012	10.727 6	2013	2014	
读者服务	2012	9.802 8	2012	2012	
著作权	2012	9.073 5	2012	2012	
科技查新	2012	9.073 5	2012	2012	
比较研究	2012	9.073 5	2012	2012	
隐性知识	2012	8.709 0	2012	2012	
元数据	2012	8.344 7	2012	2012	
数据库	2012	4.056 8	2012	2012	
学科馆员	2012	2.875 5	2012	2012	
信息资源	2012	2.254 5	2012	2012	
资源共享	2012	2.014 2	2012	2012	
信息检索	2012	1.984 2	2012	2012	
知识组织	2012	1.854 4	2013	2014	
电子政务	2012	1.577 3	2012	2014	
开放获取	2012	1.494 9	2013	2014	
评价指标	2012	1.395 3	2013	2014	

参照公式(6),将两种方式探测出的突发词分别与热点词计算覆盖率,计算结果见表 14。

观察结果发现,新模型在 3 个数据样本上的突发词覆盖率均大于 Citespace 分析结果,从而表明新模型比 Citespace 性能更好。

表 14　新模型与 Citespace 突发词探测的覆盖率

数据样本	突发时间段	新模型	Citespace
$A_{m\times i}$	2010 – 2012	12/13 = 0.92	12/20 = 0.6
$B_{m\times(i+1)}$	2011 – 2013	8/9 = 0.89	13/23 = 0.57
$C_{m\times(i+2)}$	2012 – 2014	7/10 = 0.7	9/18 = 0.5

以上分析表明本研究设计的突发词探测模型能有效发现潜在研究热点,为科研工作者把握发展趋势,捕捉研究热点提供精准服务。

5　结语

提出多测度的突发词探测及验证模型,以 2007 – 2017 年图情领域 18 种核心期刊的文献信息作为数据来源,固定 9 年为一个分析时间窗口,3 次滑动时间窗口,每个窗口又细分为标准窗口、观察窗口、表现窗口。依据相对词频、词频增长率、词频热度权重识别观察窗口内的突发词;通过 LDA 挖掘表现窗口热点主题词,并计算突发词覆盖率。结果 3 个时间窗口内的覆盖率均大于 70%,设计的模型能有效捕捉突发词,发现研究热点。本模型与 Citespace 突发词探测工具对照实验中,突发词覆盖率优于后者,说明本研究工作有价值。

本文研究还存在一些不足,也是未来研究的重点:①突发词识别条件的改进,提高突发词识别的准确度;②改进模型验证方式,现在突发词和热点词匹配的关系是一对多,未来研究将改为一对一;③运用其他方法,如 LDA2Vec、Word2Vec、Coder-autoencoder 等深度学习方法进行多热点对照分析,寻找最佳应用。

参考文献:

[1] 关鹏, 王曰芬. 基于 LDA 主题模型和生命周期理论的科学文献主题挖掘[J]. 情报学报, 2015, 34(3): 286 – 299.
[2] KLEINBERG J. Bursty and hierarchical structure in streams[J]. Data mining & knowledge discovery, 2003, 7(4): 373 – 397.
[3] 郑乐丹. 基于突发检测的我国数字图书馆研究前沿及其演进分析[J]. 图书馆论坛, 2013, 33(1): 47 – 51.
[4] CHEN C M. CitespaceII: detecting and visualizing emerging trends and transient patterns in scientific literature[J]. Journal of the Association for Information Science & Technology, 2006, 57(3): 359 – 377.
[5] 杨选辉,蔡志强. 基于突变检测与共词分析的关联数据新兴趋势探测[J]. 情报科学, 2018, 36(11): 164 – 168.
[6] 唐晓彬,周志敏,董莉. 大数据背景下网络突发事件动态监测研

究[J]. 统计研究, 2017, 34(2): 46 - 56.

[7] 卓可秋,虞为,苏新宁. 突发事件检测的 MapReduce 并行化实现[J]. 现代图书情报技术, 2015(2): 46 - 54.

[8] 陈国兰. 基于爆发词识别的微博突发事件监测方法研究[J]. 情报杂志, 2014, 33(9): 123 - 128.

[9] 逯万辉,马建霞. 基于 CRFs 的领域爆发词识别的研究与实现[J]. 情报科学, 2014, 32(1): 89 - 93.

[10] 介飞,谢飞,李磊,等. 社交网络中隐式事件突发性检测[J]. 自动化学报, 2018, 44(4): 730 - 742.

[11] XIE W, ZHU F, JIANG J, et al. TopicSketch: real-time bursty topic detection from Twitter[J]. IEEE transactions on knowledge and data engineering,2016,28(8):2216 - 2229.

[12] 王莉亚. 基于关键词突变的主题突变研究[J]. 情报理论与实践, 2013, 36(11): 45 - 48.

[13] 王征,易莉,赵磊. 基于突发词检测的科研热点发掘服务模型研究[J]. 情报杂志, 2015, 34(12): 176 - 180.

[14] 张金柱,吕品. 基于主题关联度改进的主题演变和突变分析[J]. 情报理论与实践, 2018, 41(3): 129 - 135.

[15] 姜鑫,王德庄,马海群. 关键词词频变化视角下我国“科学数据”领域研究主题演化分析[J]. 现代情报, 2018, 38(1): 141 - 146,161.

[16] SHI L, DU J P, LIANG M Y. Strm: a sparse rnn-topic model for discovering bursty topics in big data of social networks[J]. Journal of information science and engineering, 2019, 35(4): 749 - 767.

[17] 傅柱,王曰芬. 共词分析中术语收集阶段的若干问题研究[J]. 情报学报,2016,35(7):704 - 713.

[18] 刘敏娟,张学福,颜蕴. 基于词频、词量、累积词频占比的共词分析词集范围选取方法研究[J]. 图书情报工作,2016,60(23):135 - 142.

[19] Wikipedia. Long tail[EB/OL]. [2019 - 09 - 08]. https://en.wikipedia.org/wiki/Long_tail.

[20] 徐剑,黄秋月. “二八定律”在图书馆管理中的应用[J]. 中国图书馆学报, 2007(5): 106 - 108.

[21] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of machine learning research, 2003, 3(4/5): 993 - 1022.

[22] 王建. 基于多特征融合的微博突发事件检测方法研究[D]. 北京:北京信息科技大学, 2018.

[23] 马文建. 基于突发词检测的中文专利预警系统[D]. 北京:北京工业大学, 2016.

[24] 安璐,杜廷尧,李纲,等. 突发公共卫生事件利益相关者在社交媒体中的关注点及演化模式[J]. 情报学报, 2018, 37(4):394 - 405.

作者贡献说明:

奉国和:论文主题思想和论文修改意见提出,论文最终版修订;

武佳佳:数据筛选和处理,实验,初稿撰写;

莫幸清:稿件校对与数据核对。

Research on Detection and Verification of Burst Words with Multiple Measures

Feng Guohe　Wu Jiajia　Mo Xingqing
Information Managment Department, School of Economics & Management,
South China Normal University, Guangzhou 510006

Abstract: [Purpose/significance] In order to effectively detect potential research hotspots in scientific and technological literature, to study the characteristic conditions of keyword emergencies in the literature, and to construct a model of burst word recognition is of great significance to promote scientific researchers to accurately grasp the research direction. [Method/process] This paper got keywords and word frequency in each year, constructed keyword-year matrix, divided the analysis period into standard window, observation window and performance window, used multi-measure burst word detection model to identify keywords with burst characteristics in the observation window, and used LDA to mine topic words as hot words set in the performance window. The coverage index of burst words was designed, and the sliding time window method was used to calculate the coverage of burst words and hot words in different time windows to verify the accuracy of model recognition. [Result/conclusion] The three sliding time windows calculated that the coverage of the three sudden words is more than 70%. In the control test with Citespace, the coverage of the model three times is greater than the former, indicating that the designed burst word detection model performs well.

Keywords: burst word detection　sliding time window　multiple measures　LDA topic mining